Abdominal Medical Images Segmentation with CNN-Transformer Hybrid Model

EECS 545 Project Report

Chuoqi Chen chuoqic@umich.edu
Haowen Chen haowen@umich.edu
Yuang Lu yuanglu@umich.edu
Yuqi Yan yukei@umich.edu
Ying Yuan yyyuan@umich.edu

1 Introduction

Medical image segmentation plays a crucial part in identifying pixels of anatomical objects from medical images during computer-aided diagnosis, treatment planning, and computer-integrated surgery. However, conducting segmentation on CT images by hand is a time-consuming task with low inter-rater agreement for radiologists and doctors(Liu et al., 2020). Therefore, it's necessary to apply new technologies that can accelerate segmentation on CT images.

The primary challenges for medical image segmentation mainly lie in three aspects: (1) Complex boundary interactions, (2) Large appearance variation, (3) Low tissue contrast(Zhou et al., 2019). Conventionally, meaningful or task-related features were mostly designed by human experts based on their knowledge about the target domains(Shen et al., 2017). These challenges lead to complex features that are hard to depict. With the advance in deep learning, people find out application of convolutional neural network shows an edge in feature representation of medical image. Since then, many deep learning models have been applied to this significant work(Wang et al., 2022; Minaee et al., 2021). However, new challenges emerge when applying deep learning techniques, like over-fitting due to the scarcity of the training dataset, extensive training time, variable scales among target objects(Hesamian et al., 2019).

In our study, we mainly focus on the abdominal multi-organs segmentation problem. We purpose two novel neural network structures to solve two problems that caught our attention when we do our preliminary researches: fragmental object edges and failure to detect small, direction-dependent objects. We expect to enhance medical image segmentation by integrating Self Attention mechanism into convolutional neural network.

2 Related work

In this section, we introduce recent work for medical image segmentation. The objective of image semantic segmentation is to apply pixel-wise classification for the images. To achieve this, researchers proposed the encoder-decoder network structure. Among these networks, UNet(Ronneberger et al., 2015) is one of the most popular end-to-end structure used for medical image segmentation. There have been many advancements in UNet architecture by researchers implementing new methods or incorporating other methods into UNet.

2.1 ResUNet

Residual UNet(Zhang et al., 2018), also known as ResUNet, is inspired by the ResNet(He et al., 2016) structure. Previous experiments have shown that the increasing number of network layers would lead to loss of feature identities and degradation of performance. Such degradation is caused by vanishing gradients in deeper network. The skip connections before downsampling or upsampling in ResUNet help to preserve feature maps by concatenating them from previous layers to deeper layers, thereby alleviate the vanishing gradient problem and allow deeper networks to be designed.

ResUNet is capable of detecting all 13 organs in the Synapse dataset including small and direction-dependent objects such as left and right adrenal gland while all other transformer-based models failed to do so. It can also perform better in segmenting other small objects such as portal vein and splenic vein. We believe it is ResUNet's strong ability of capturing low-level detail with the help of convolutional layer that makes the model possible to classify these small and direction-dependent objects.

2.2 TransUNet

TranUNet(Chen et al., 2021) combines Transformers and U-Net together, leveraging both detailed high-resolution spatial information from CNN features and the global context encoded by Transformers. Transformer serves as a strong encoder, providing attentive feature sequence. The decoder upsamples the encoded features which are then combined with different high-resolution CNN features skipped from the encoding path. Empirical results show that this architecture achieves superior performances to various competing methods on medical image segmentation tasks.

Since it's the first medical image segmentation framework that fully leverage the power of Transformers into U-Net, we consider TransUNet as the baseline model in this project, and implement a novel model in section 3.

Hybrid Convolution Neural Network (CNN)-Transformer architecture is employed as encoder part of the TransUNet. CNN is utilized as a feature extractor to generate detailed high-resolution spatial feature map as the input for the Transformer. The Transformer tokenizes low-level CNN features as sequences to extract high-level global context. Detailed implementation in Transformer is as below:

Image-to-Sequence Generator converts images to long sequences as the input of Transformers. Given 3D input volume $x \in \mathbb{R}^{H\times W\times D\times C}$ with resolution (H, W, D) and C input channels, the generator converts x to flattened uniform non-overlapping $P \times P \times P$ patches:

$$x \in \mathbb{R}^{H\times W\times D\times C} \rightarrow x_s \in \mathbb{R}^{N\times(P^3\cdot C)}$$

where $N = (H \times W \times D)/P^3$ is the length of the sequence. Then the sequence is mapped to a K dimensional embedding space by

$$z_s = \left[x_s^1 E; x_s^2 E, ..., x_s^N E\right] + E_{pos}$$

where $E \in \mathbb{R}^{P^3\cdot C\times K}$, and $E_{pos} \in \mathbb{R}^{N\times K}$ is learnable 1D embedding.

Multi-head Self Attention is widely used to solve NLP problems. A MSA layer consists of n parallel SA heads. Each SA head is a parameterized function that learns the mapping between a query($\boldsymbol{q}$) and the corresponding key($\boldsymbol{k}$)-value($\boldsymbol{v}$) representations in $\mathbb{R}^{N\times K}$. Attention weights A are computed by measuring the similarity between two elements in z and their key-value pairs:

$$A = softmax(\frac{\boldsymbol{q}\boldsymbol{k}^T}{\sqrt{K/n}})$$

Then we have the output

$$SA(z) = A\boldsymbol{v}$$
$$MSA(z) = [SA_1(z), ..., SA_N(z)]W_{msa}$$

where $W_{msa} \in \mathbb{R}^{K\times K}$ is trainable weights.

Transformers block is implemented based on the Multi-Head Self Attention mechanism. Denote $z_{i-1} \in \mathbb{R}^{N\times K}$ the input of the $i-1$-th transformers layer. For the i-th layer,

$$z_i' = MSA(Norm(z_{i-1})) + z_{i-1}$$

where $i = 0, ..., L$. The output is then normalized, connected to a full connected layer and a residual block.

$$z_i = MLP(Norm(z_i')) + z_i'$$

Cascaded Upsampler (CUP) is employed as the decoder in TransUNet to upsample the high-level semantic features and decode upsampled features combined with the high-resolution CNN features to enable precise localization and generate segmentation outcome.

2.3 Swin-UNet

Based on the outstanding performance in TransUNet, it is possible to bring optimization of transformer to image segmentation. Swin-UNet could be a possible solution. Like the Trans-UNet, Swin-UNet also segmentates image in small pieces as the input of transformers. The difference is that the slice in the image is connected to the rest other slices in Trans-UNet model, while the limited slices is just connected with the neighbor slices into one transformer. In the higher layer, the output from the neighbor transformer will be the input of the next layer transformer. By this way, more details from the images could be trained while neighbor information could affects each other. (Cao et al., 2021)

Though Swin-UNet has overall higher DICE score than TransUnet, meaning that it can classify each pixel in the right class better than TransUNet, we found that the segmented objects in the image generated by Swin-UNet is fragmental and have rougher edges compared to that of TransUNet.

2.4 UTNet

Although transformer variants show powerful performance in vision tasks, pixels packed in the same input image patch does not contain enough spatial information, which lead to weaker inductive bias compared with traditional CNN. Therefore, vision transformer-based models require large image set to gain enough prior knowledge in image domain. However, a major characteristic of medical image segmentation tasks is that they usually cannot provide such a big dataset, using pretrained weight on other large image datasets then becomes a widely adopted solution.

UTNet, unlike other models, employs another method. Unlike Common vision transformer and their variants packing pixel into patches and flattening them into vectors, and UTNet preserves the first few encoding layer as convolutional layer, the pixel-level spatial relations thus can be retained (Gao et al., 2021).

2.5 UNETR

UNETR (Hatamizadeh et al., 2021) proposed a novel transformer-based model for volumetric medical image segmentation. Instead of extracting features with Transformers at the bottleneck of U-Net, UNETR has its skip-connected decoder combines representations from multiple intermediate Transformers layers. Deconvolutional layers are applied to these intermediate representations to increase the resolution. Such architecture helps the network capture global contextual representation at multiple scales and increase the capability for learning long-range dependencies.

2.6 TransFuse

To improve the efficiency of modeling global contexts while preserving a solid grasp of low-level features is needed for the segmentation task, TransFuse uses a parallel approach to merge Transformers and CNNs. This allows the efficient collection of both global dependencies and low-level spatial features in a significantly shallower manner. The BiFusion module is a novel fusion technique that efficiently fuses the multi-level features from both branches. TransFuse obtains the state-of-the-art results on both 2D and 3D medical picture sets with significantly fewer parameters and significantly faster inference(Zhang et al., 2021).

3 Proposed Method

The primary model for improvements is TransFuse. We implement the original TransFuse model as one of the baselines, and propose novel models by modifying TransFuse with the stack of transformer blocks from Swin-Transformer and UTNet as the encoder to learn sequence representations the input and further enhance the efficiency in capturing the global information. The overall diagram of the proposed model is shown in Fig. 1.

This model consists of two parallel branches that process information differently:

(1) CNN Branch

It gradually increases the receptive field and encodes the feature from local to global. With the Transformer branch to obtain global context information instead, CNN branch could be shallower, and the proposed model could be also retaining richer local information. We fuse the output of the 3rd block ($\mathbf{l}^1 \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times C_1}$), 2nd block ($\mathbf{l}^2 \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times C_2}$), and 1st block ($\mathbf{l}^3 \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times C_3}$)) with the results from Transformer. We propose different models with CNN variants. Please see section 3.1 and 3.2 for details.

(2) Transformer Branch

It starts with global self-attention and recovers the local details at the end. The input image $\mathbf{x} \in \mathbb{R}^{H\times W\times 3}$ is first evenly divided into $N = \frac{H}{16}\times\frac{W}{16}$ patches. Then we use the consecutive $3\times 3\times 3$ upsampling-convolution layers to obtain $\mathbf{z}^1 \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times D_1}$, $\mathbf{z}^2 \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times D_2}$, and $\mathbf{z}^3 \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times D_3}$. We propose different models with transformer variants. Please see section 3.1 and 3.2 for details.

(3) BiFusion Module

It fuses the encoded features of the same resolution extracted from CNN and Transformer branch. The fused feature $\mathbf{b}^i, i = 1, 2, 3$ is obtained by the following operations:

$$\hat{\mathbf{z}}^i = \text{ChannelAttn}(\mathbf{z}^i)$$

$$\hat{\mathbf{l}}^i = \text{SpatialAttn}(\mathbf{l}^i)$$

$$\hat{\mathbf{c}}^i = \text{Conv}(\mathbf{z}^i\mathbf{W}_1^i \odot \mathbf{l}^i\mathbf{W}_2^i)$$

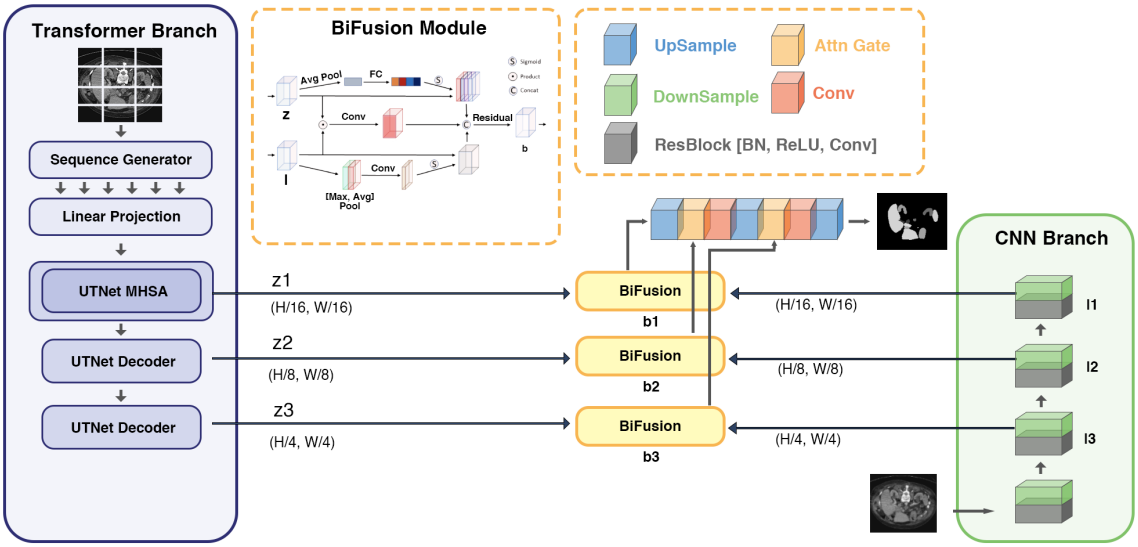

Figure 1: Network Structure of UT-Fuse

$$\hat{\mathbf{b}}^i = \text{Residual}([\hat{\mathbf{c}}^i, \hat{\mathbf{z}}^i, \hat{\mathbf{l}}^i])$$

, where $\mathbf{W}_1^i \in \mathbb{R}^{D_i \times L_i}$, $\mathbf{W}_2^i \in \mathbb{R}^{C_i \times L_i}$, $|\odot|$ is the Hadamard product and Conv is a 3×3 convolution layer. The channel attention is implemented according to the SE-Block proposed by Hu et al. (2018) to facilitate global information from the Transformer branch. The spatial attention is adopted from CBAM (Woo et al., 2018) block as spatial filters to enhance local details.

(4) Attention-gated skip-connection

It combines fused feature maps and generates the segmentation(Schlemper et al., 2019). We obtain

$$\hat{\mathbf{f}}^{i+1} = \text{Conv}([\text{Up}(\hat{\mathbf{f}}^i), \text{AG}(\mathbf{f}^{i+1}, \text{Up}(\hat{\mathbf{f}}^i))])$$

and $\mathbf{f}^1 = \hat{\mathbf{f}}^1$.

With this branch-in-parallel approach, firstly, we can capture global information and preserve sensitivity on low-level context without building deep nets; secondly, the fused representation would be powerful and compact with BiFusion module by utilizing the different features of CNN and Transformer simultaneously.

3.1 Swin-Fuse

In order to solve the problem of fragmental edges of Swin-UNet, we apply the BiFusion module to the Swin-Transformer. By fusing the encoder of ResNet and encoder of Swin-Transformer together, low level details captured by the CNN branch and the global information captured by the transformer branch can be combined together by the BiFusion module and then sent to the decoder. The hybrid model therefore can preserve the strength of high prediction accuracy of Swin-UNet and further improve the model's ability of restoring edges of organs with the help of the captured low-level details.

Swin-Fuse is trained with ADAM optimizer, with learning rate set to 0.001. Beta1 and beta2 are set to 0.5 and 0.999 correspondingly. The loss of the output is the weighted sum of the cross entropy and DICE score with weights of 0.4 and o.6. The cross entropy and DICE loss are calculated based on the intermediate result and the final result, which is consistent to the method introduced by the TransFuse.

3.2 UT-Fuse

In practice, all transfomer-based models such as Trans-UNet, Swin-UNet, UTNet and hybrid model like TransFuse and Swin-Fuse failed to segment left adrenal gland and right adrenal gland. After analyzing the characteristics of these two organs, we found that they are small and direction-dependent organs, that they are located almost in the same slice of the 3D CT scan, and only differentiating in the direction (left and right oriented) and the

relative location in the 2D image. In order to successfully segment these two organs, we need to enhance the model's ability of classifying small organs based on the neighboring organ and also tell the direction of organs.

Therefore, we propose UT-Fuse, fused by ResNet and UTNet with the BiFusion module. The model is capable of capturing pixel-level details from the input image from both of the transformer branch and the CNN branch. The transformer branch will also process the location information with global self-attention. Since both branches have intermediate convolutional layer, neighboring organs have stronger connection and the model can detect direction dependent organs based on the relative location with other organs.

Training method and parameters of UT-Fuse is consistent with that of Swin-Fuse

4 Implementation

In this project, we use Synapse multi-organ image segmentation dataset (Sage Bionetworks) that was originally used for the MICCAI 2015 Multi-Atlas Abdomen Labeling Challenge. 50 abdomen CT scans are included in the dataset. To align with previous works, 30 out of 50 scans are used for this project. These CT scans range from $512 \times 512 \times 85$ to $512 \times 512 \times 198$ in pixels with resolution varying from $0.54 \times 0.54\ mm^2$ to $0.98 \times 0.98\ mm^2$. 13 different organs, including spleen, stomach, liver, etc. are labeled pixel by pixel.

We cut the 3D CT scan into 512×512 pixel 2D images, and split 12 out of 30 scans as testing image, which is consistent to preprocessing techniques mentioned in SwinUNet and TranUNet. Models such as TranUNet and SwinUNet use pretrained weight on large image dataset while novel architectural models like UTNet use unique downsampling method that requires specific image dimension. Therefore, we crop images to [224,224] or [256, 256] in dimension corresponding to the model requirements. In order to improve the robustness of the model, data augmentation methods such as random flip and random rotation have been applied.

5 Evaluation

Though many transformer-based model reported better performance compared to traditional ResUNet, we only use models of comparable sizes for evaluation. For example, Res-UNet based on ResNet-34 which has 63.5 million parameters and TransUNet based on vit-base-patch16-224 which has 86 million parameters are used as reference model. Our implemented hybrid model use pre-trained weight of ResNet-34 and Tiny Swin-Transformer (28 million parameters).

For medical images segmentation, there are two types of errors related to segmentation accuracy, namely the delineation of the boundary (contour) and the size (volume of the segmented object)(Taha and Hanbury, 2015). Here we select two metrics for these two types of errors based on literature review.

Dice Coefficient, also called the overlap index, is the most used metric in validating medical image segmentation, which is defined as

$$DICE = \frac{2TP}{2TP + FP + FN}$$

Hausdorff Distance(HD) measures the distance between crisp volumes. Given two point sets A and B, HD is defined by

$$HD(A, B) = \max(h(A, B), h(B, A))$$

where $h(A, B)$ is called the directed Hausdorff distance and given by

$$h(A, B) = \max_{a \in A} \min_{b \in B} ||a - b||$$

where $||\cdot||$ is a norm(normally Euclidean distance). Note that HD is sensitive to outliers and we may use the q-th quantile of distances instead of the maximum, so that possible outliers are excluded.

6 Results Analysis

We report the average Dice Coefficient and average Hausdorff Distance (HD) on 13 abdominal organs (spleen, right kidney, left kidney, gallbladder, esophagus, liver, stomach, aorta, inferior vena cava, portal vein and splenic vein, pancreas, right adrenal gland, and left adrenal gland) with a random split of 18 training cases (2212 axial slices) and 12 cases for validation. Check table 1 and 2 for results comparison.

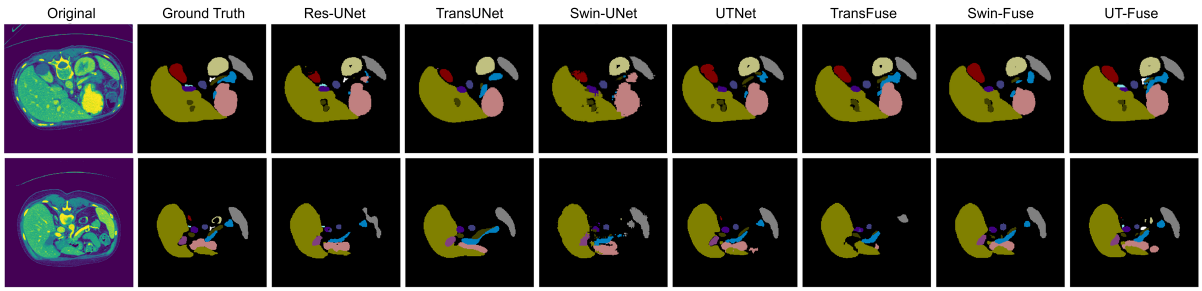

Figure 2: Segmentation result of Synapse dataset.

DICE Score	ResUNet	TransUNet	SwinUNet	UTNet	TransFuse	Swin-Fuse	UT-Fuse
spleen	81.43%	83.43%	87.49%	86.46%	86.07%	89.45%	87.10%
right kidney	71.77%	69.89%	77.00%	83.03%	79.07%	79.81%	72.90%
left kidney	80.12%	74.83%	81.15%	86.12%	79.08%	86.64%	82.11%
gallbladder	61.17%	45.57%	60.04%	70.97%	26.68%	63.96%	58.67%
esophagus	71.53%	54.54%	67.64%	73.76%	69.90%	66.48%	66.12%
liver	93.71%	91.51%	93.42%	95.42%	93.79%	93.21%	93.92%
stomach	73.92%	72.43%	72.46%	75.50%	71.45%	73.36%	80.70%
aorta	87.39%	70.97%	83.18%	89.99%	85.94%	86.34%	87.34%
inferior vena cava	74.47%	59.69%	70.57%	81.33%	76.52%	77.59%	69.65%
portal vein and splenic vein	66.32%	39.60%	54.13%	66.69%	55.67%	57.43%	62.08%
pancreas	61.22%	45.00%	54.81%	62.34%	62.30%	56.49%	61.72%
right adrenal gland	59.26%	0.00%	0.00%	0.00%	0.00%	0.00%	45.91%
left adrenal gland	58.21%	0.00%	0.00%	0.00%	0.00%	0.00%	50.33%
Mean	72.35%	54.42%	61.68%	67.05%	60.50%	63.90%	70.66%

Table 1: Mean_dice by organs

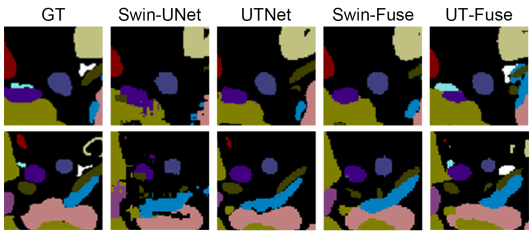

Figure 3: Local segmentation result for small objects

6.1 Swin-Fuse

With the help of BiFusion module and CNN-branch, Swin-Fuse has better capability to capture the local interactions with neighboring pixels and therefore can restore the edge of detected objects better. Intuitively, the classified organs of the Swin-Fuse's segmented output is properly shaped, while that of Swin-UNet is fragmental. as showed in Fig.2 Numerically, Swin-Fuse has better mean DICE score and mean HD95 (63.90%, 18.22) compared with TransUNet (54.42%, 23.54), SwinUNet (61.68%, 22.82), and TransFuse (60.50%, 18.44).

6.2 UT-Fuse

UT-Fuse which fused by hybrid UTNet and ResNet-34 has a better DICE score compared to other Transformer-based models. With a margin of 3.61% higher than UTNet (67.05%) and 16.24% higher than the original TransUNet. The HD95 of UT-Fuse is also better than other models. The mean DICE score of UT-Fuse is 15.73, betting the second place, Swin-Fuse, by 18.22. UT-Net can also classify and segment small and direction dependent objects such as right and left adrenal gland which all other Transformer-based models failed to segment, with DICE score of 45.91% and 50.33%, and HD95 of 7.55 and 6.40. Though UT-Fuse has lower DICE score than Res-UNet (70.66% compared to 72.35%), the HD95 of UT-Fuse (15.73) is much better than that of Res-UNet (24.92).

The performance of UT-Fuse is generally better than almost all other models, however, it is inferior to UTNet in several classes. An interesting observation is that both Res-UNet and UTNet have higher DICE score for gallbladder than UT-Fuse even though UT-Fuse is composed of these two

HD95	ResUNet	TransUNet	SwinUNet	UTNet	TransFuse	Swin-Fuse	UT-Fuse
spleen	68.37	31.07	29.63	35.77	34.35	33.60	8.78
right kidney	74.55	62.91	31.54	54.66	41.48	19.52	39.99
left kidney	39.32	40.32	47.11	35.62	22.19	16.18	39.22
gallbladder	24.36	38.71	29.26	10.95	10.21	10.85	14.52
esophagus	6.52	7.94	5.93	5.14	4.61	8.78	6.11
liver	26.54	25.21	21.05	12.77	14.70	27.26	20.18
stomach	23.10	17.44	18.28	17.64	25.85	16.32	11.90
aorta	15.49	11.59	10.69	4.27	10.96	11.90	6.35
inferior vena cava	8.30	18.33	12.89	5.43	9.28	6.16	8.86
portal vein and splenic vein	15.50	32.08	29.13	17.45	16.72	29.02	23.51
pancreas	10.63	20.35	15.48	12.62	12.45	20.80	11.13
right adrenal gland	5.31	0.00	0.00	0.00	0.00	0.00	7.55
left adrenal gland	5.94	0.00	0.00	0.00	0.00	0.00	6.40
Mean	24.92	23.54	22.82	19.30	18.44	18.22	15.73

Table 2: Mean_hd95 by organs

models. This problem is possibly caused by unclear boundary of gallbladder with other organs, when the BiFusion module fusing two separate models, information captured by lower layers are mixed up, which makes the model harder to separate Gallbladder and other organs.

7 Contribution

All co-authors equally contributed to this project. All co-authors discussed the design of the novel network structure and experiments and involved in writing the report.

Yuqi Yan design and implemented UT-Fuse model, performed experiments with ResUNet, preprocessed Synapse dataset and created figures for the report;

Yuang Lu design and implemented Swin-Fuse model, trained model for TransUNet and organized all testing results;

Chuoqi Chen design and implemented Swin-Fuse model, set up experimental environment of Swin-UNet model, trained and validated SwinUNet on Synapse dataset;

Haowen Chen design and implemented UT-Fuse model, migrated the UTNet model from MM dataset to Synapse dataset, trained the model and gathered experimental result;

Ying Yuan design and implemented UT-Fuse model, set up the experimental environment of TransFuse model, pre-processed ISIC2017 data, migrated TransFuse model to Synapse, trained TransFuse on Synapse dataset, trained and tested TransFuse on ISIC2017 dataset.

References

Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. 2021. Swin-unet: Unet-like pure transformer for medical image segmentation.

Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. 2021. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*.

Yunhe Gao, Mu Zhou, and Dimitris N Metaxas. 2021. Utnet: a hybrid transformer architecture for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 61–71. Springer.

Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger Roth, and Daguang Xu. 2021. Unetr: Transformers for 3d medical image segmentation.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Mohammad Hesam Hesamian, Wenjing Jia, Xiangjian He, and Paul Kennedy. 2019. Deep learning techniques for medical image segmentation: achievements and challenges. *Journal of digital imaging*, 32(4):582–596.

Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141.

Liangliang Liu, Jianhong Cheng, Quan Quan, Fang-Xiang Wu, Yu-Ping Wang, and Jianxin Wang. 2020.

A survey on u-shaped networks in medical image segmentations. *Neurocomputing*, 409:244–258.

Shervin Minaee, Yuri Y Boykov, Fatih Porikli, Antonio J Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. 2021. Image segmentation using deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

info@sagebase.org Sage Bionetworks. Sage bionetworks.

Jo Schlemper, Ozan Oktay, Michiel Schaap, Mattias Heinrich, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. 2019. Attention gated networks: Learning to leverage salient regions in medical images. *Medical image analysis*, 53:197–207.

Dinggang Shen, Guorong Wu, and Heung-Il Suk. 2017. Deep learning in medical image analysis. *Annual review of biomedical engineering*, 19:221–248.

Abdel Aziz Taha and Allan Hanbury. 2015. Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. *BMC medical imaging*, 15(1):1–28.

Risheng Wang, Tao Lei, Ruixia Cui, Bingtao Zhang, Hongying Meng, and Asoke K. Nandi. 2022. Medical image segmentation using deep learning: A survey. *IET Image Processing*.

Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. 2018. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19.

Yundong Zhang, Huiye Liu, and Qiang Hu. 2021. Transfuse: Fusing transformers and cnns for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 14–24. Springer.

Zhengxin Zhang, Qingjie Liu, and Yunhong Wang. 2018. Road extraction by deep residual u-net. *IEEE Geoscience and Remote Sensing Letters*, 15(5):749–753.

Sihang Zhou, Dong Nie, Ehsan Adeli, Jianping Yin, Jun Lian, and Dinggang Shen. 2019. High-resolution encoder–decoder networks for low-contrast medical image segmentation. *IEEE Transactions on Image Processing*, 29:461–475.